

UNIVERSITY OF SHARJAH
TUTORIALS 3
EXPLORATORY DATA ANALYSIS

AHMED HOSSAIN, PhD

Exploratory Data Analysis

Biostatistics

DATA SUMMARIES

- Tabular: Frequencies, relative frequencies etc.
- Graphical: Line graph/ diagram, Bar charts/ plot, histograms, scatter plots, box plots, pie chart etc.

Biostatistics for categorical data

DATA SUMMARIES

FREQUENCY TABLES Frequency Tables used to summarize

- Nominal or ordinal data having natural categories
- Discrete or continuous data, usually after data have been grouped into categories

```
. tabulate gender
```

gender	Freq.	Percent	Cum.
Female	965	57.07	57.07
Male	726	42.93	100.00
Total	1,691	100.00	

```
. tabulate smoke
```

smoke	Freq.	Percent	Cum.
No	1,270	75.10	75.10
Yes	421	24.90	100.00
Total	1,691	100.00	

Biostatistics for categorical data

DATA SUMMARIES: BIVARIATE TABLE

```
. tabulate gender smoke, row
```

Key	
frequency	row percentage

gender	smoke		Total
	No	Yes	
Female	731	234	965
	75.75	24.25	100.00
Male	539	187	726
	74.24	25.76	100.00
Total	1,270	421	1,691
	75.10	24.90	100.00

Biostatistics for categorical data

DATA SUMMARIES: LINE GRAPH/ DIAGRAM

- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables.

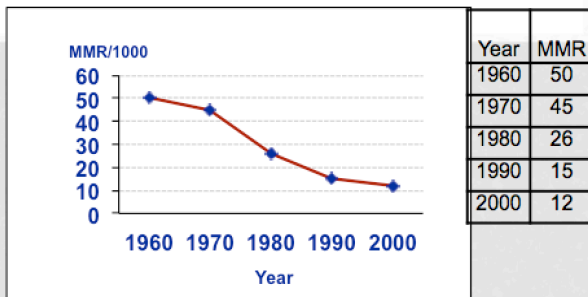
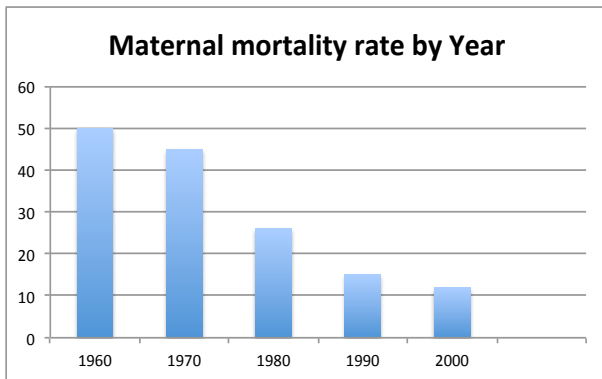


Figure: Maternal mortality rate of (country),
1960-2000

Biostatistics for categorical data

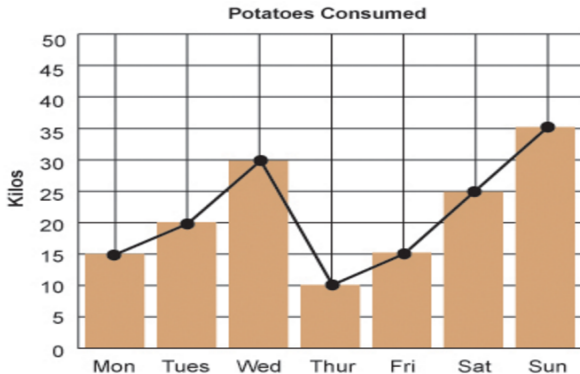
DATA SUMMARIES: BAR CHART/ DIAGRAM

- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables.



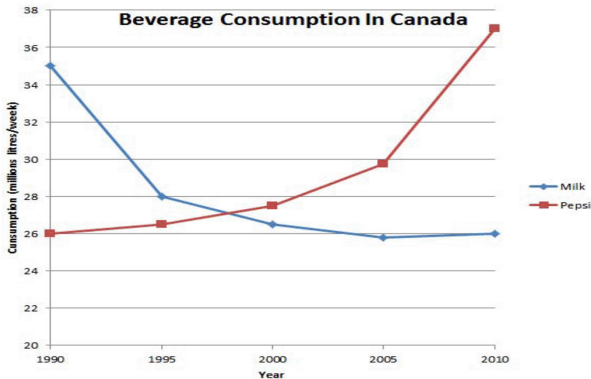
Biostatistics for categorical data

DATA SUMMARIES: BAR CHART AND LINE GRAPH



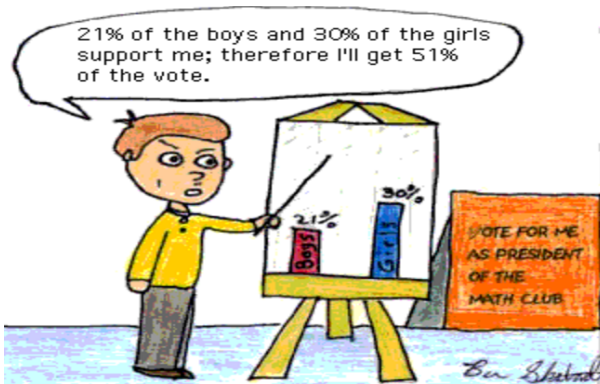
Biostatistics for categorical data

DATA SUMMARIES: LINE GRAPH



Interpreting data correctly is Important.

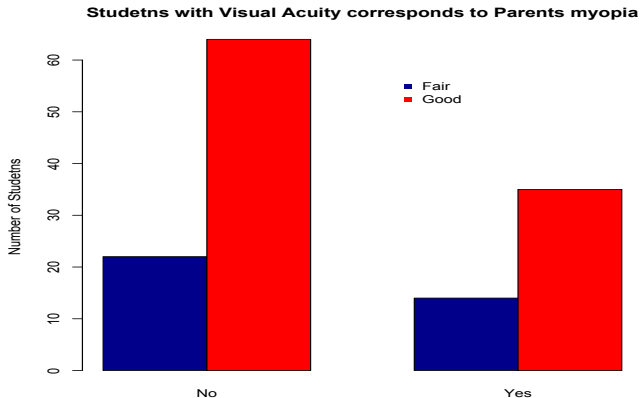
WHAT IS WRONG HERE?



Biostatistics for categorical data

DATA SUMMARIES: BARPLOT

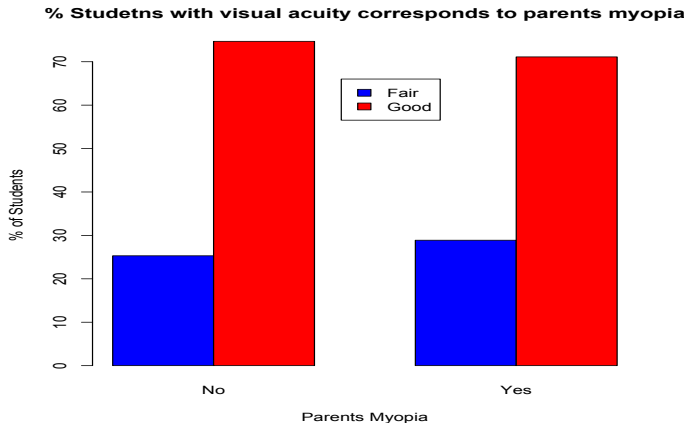
- Find a limitation of this barplot. It is in terms of interpretation.



Biostatistics for categorical data

DATA SUMMARIES: BARPLOT

Always display bar graphs with percentages.

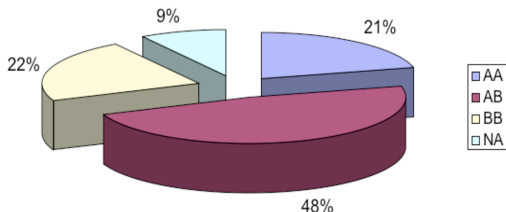


Biostatistics for categorical data

DATA SUMMARIES: PIE CHART

- Used to express information from frequency summary table of categorical data.
- Circle divided into slices- number of slices corresponds to the number of categories
- Relative frequency percent make it easier to create a proportional pie chart.

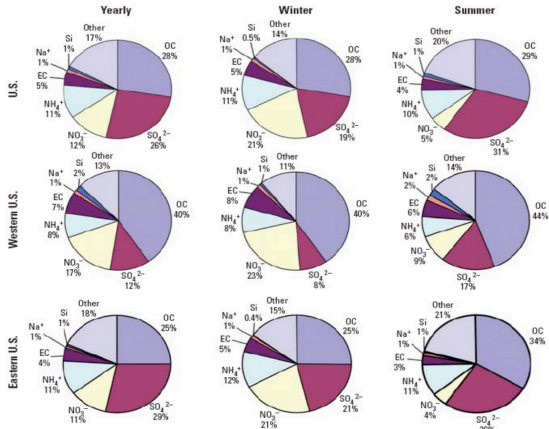
Distribution of genotypes



Biostatistics for categorical data

DATA SUMMARIES: LIMITATIONS OF PIE CHARTS

It is hard to follow the data summaries with pie charts when a categorical variable has many categories or bi-variate table provides many categories.



For quantitative (discrete or continuous) data


STEM-AND-LEAF PLOTS (STEMPLOTS)

- Used to visualize distribution (shape, center, range, variation) of continuous variables and small data.

05 11 21 24 27 28 30 42 50 52

- Plot all data points and rearrange in **rank order**:

```
0 | 5
1 | 1
2 | 1478
3 | 0
4 | 2
5 | 02
x10
```

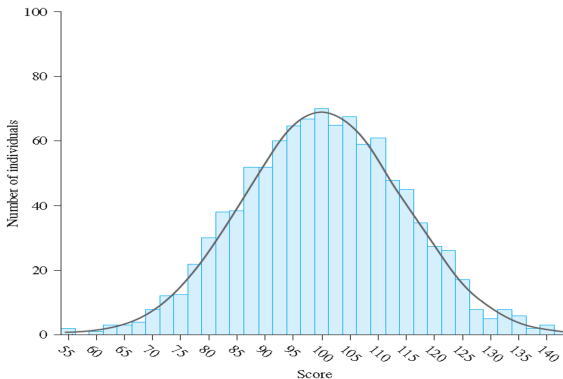
- Here is the plot **horizontally**:  (for demonstration purposes)

```
      8
      7
      4      2
5 1 1 0 2 0
-----
0 1 2 3 4 5
-----
Rotated stemplot
```

For quantitative (discrete or continuous) data

DATA SUMMARIES: HISTOGRAM

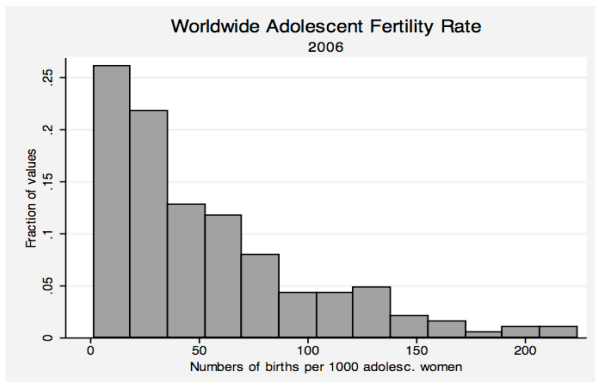
- Used to visualize distribution (shape, center, range, variation) of continuous variables and large data.
- “Bin size” is important.



For quantitative (discrete or continuous) data

DATA SUMMARIES: HISTOGRAM

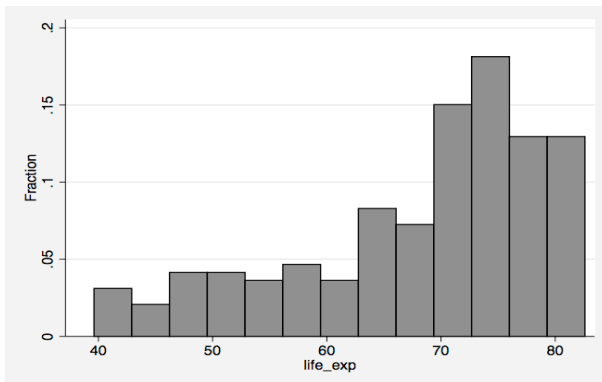
Positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right.



For quantitative (discrete or continuous) data

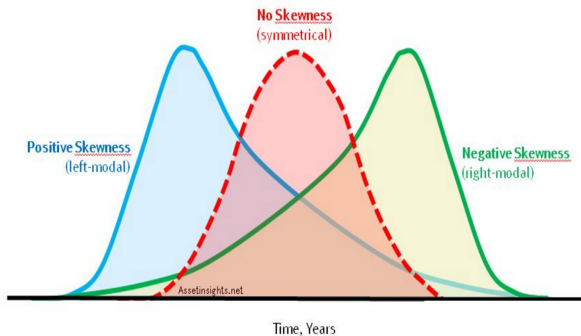
DATA SUMMARIES: HISTOGRAM

Negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left.



For quantitative (discrete or continuous) data

DATA SUMMARIES: SKEWNESS



For quantitative (discrete or continuous) data

EFFECT OF BIN SIZE ON HISTOGRAM

Length of time of service calls at a bank

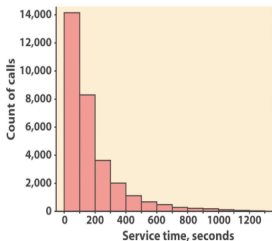


Figure 1-4
Introduction to the Practice of Statistics, 9th Edition
© 2009 W. H. Freeman and Company

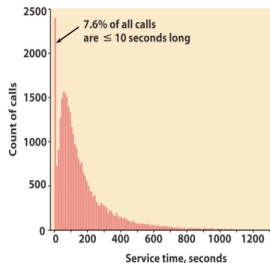


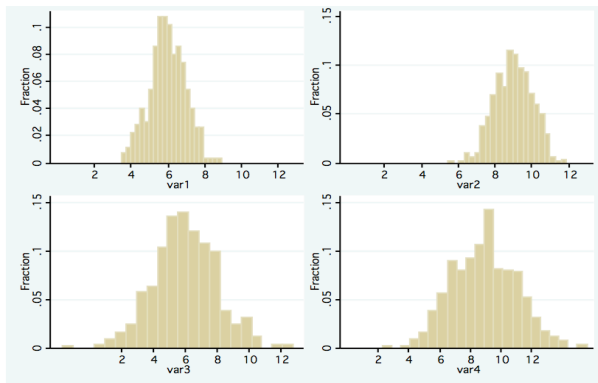
Figure 1-2
Introduction to the Practice of Statistics, 9th Edition
© 2009 W. H. Freeman and Company

For quantitative (discrete or continuous) data

DATA SUMMARIES: LOCATION AND SHAPE

MEASURES OF CENTRAL TENDENCY Mean, Median and Mode.

MEASURES OF SPREAD Range, Interquartile range, variance and standard deviation.



For quantitative (discrete or continuous) data

CENTRAL LOCATION: MEAN AND MEADIAN

MEAN :

To calculate the average \bar{x} of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

MEDIAN is the exact middle value.

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them

Example

Some data:

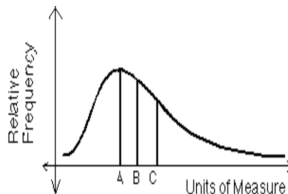
Age of participants: 17 19 21 22 23 23 23 38

$$\text{Median} = (22+23)/2 = 22.5$$

For quantitative (discrete or continuous) data

QUESTION

Which of the following orders correctly represents the measures of central tendency for the distribution shown here?

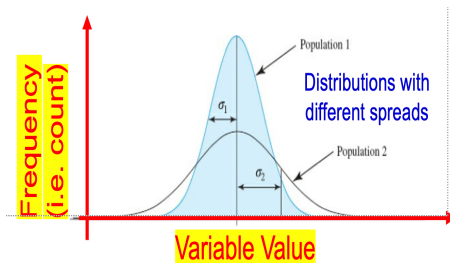


- a. A: mean, B: median, C: mode
- b. A: mode, B: mean, C: median
- c. A: median, B: mode, C: mean
- d. A: median, B: mean, C: mode
- e. A: mode, B: median, C: mean
- f. None of these orders are correct.

For quantitative (discrete or continuous) data

SPREAD: VARIANCE AND STANDARD DEVIATION

- The term spread is an informal way to refer to the **dispersion or variability** of data points. The following Figure shows distributions with different variability.
- Populations 1 and 2 have the same central locations, but population 2 has greater spread (variability).



For quantitative (discrete or continuous) data

SPREAD: VARIANCE AND STANDARD DEVIATION

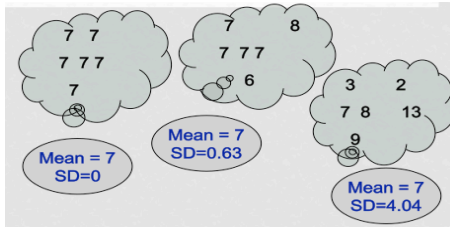
VARIANCE Average of squared deviations of values from the mean.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- Increasing contribution to the variance as you go farther from the mean.

STANDARD DEVIATION Standard deviations are simply the square root of the variance.

- Roughly 68% of the observations in the list of data lie within 1 standard deviation of the average.
- 95% of the observations lie within 2 standard deviations of the average.



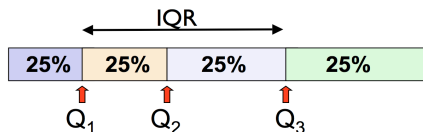
For quantitative (discrete or continuous) data

VARIANCE: WHICH ONE HAS LESS STANDARD DEVIATION?



For quantitative (discrete or continuous) data

SPREAD: QUARTILES AND INTER QUARTILE RANGE



Q1 The first quartile, Q_1 , is the value for which 25% of the observations are smaller and 75% are larger.

Q2 Q_2 is the same as the median (50% are smaller, 50% are larger)

Q3 Only 25% of the observations are greater than the third quartile.

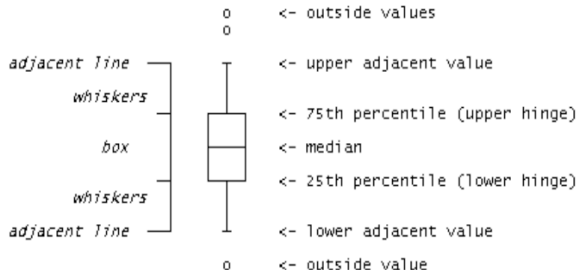
IQR It is the difference between third and first quartile.

EXAMPLE Graduate student ages: 27, 28, 31, 35, 35, 40, 42, 43, 50, 52.

- $P_{50} = Q_2$ = average of the middle two observations = $(35+40)/2 = 37.5$ years.
- $P_{25} = Q_1$ = middle observation of the lower 5 observations = 31 years.
- $P_{75} = Q_3$ = middle observation of the upper 5 observations = 43 years.

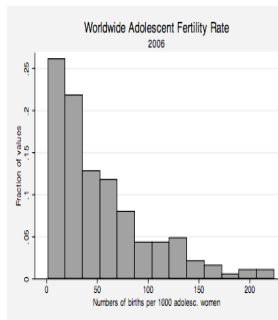
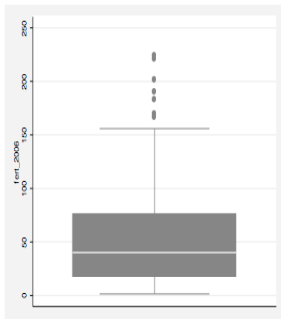
For quantitative (discrete or continuous) data

DATA SUMMARIES: BOX PLOT



For quantitative (discrete or continuous) data

DATA SUMMARIES: BOX PLOT



For quantitative (discrete or continuous) data

QUESTION: COMPARE THE BOX PLOTS?

